EUPORIAS seasonal wind prediction prototype: User evaluation

Dr. Stephann Makri, City University London (Stephann@city.ac.uk) November 2015

Executive summary

We conducted a formative user evaluation of version 2.1.0. of the EUPORIAS seasonal wind prediction prototype with 5 potential users from the wind energy industry. We asked them to explore the prototype while thinking-aloud performing realistic tasks and asked them questions aimed at probing their understanding of the prototype – at multiple 'layers.' We also asked them questions about how useful and easy to use they thought the prototype was and how likely they were to use the prototype in the future. We identified a broad range of potential design improvements. The potential users found it difficult to comment on usefulness or likeliness of use as they found it difficult to separate the prototype tool (which they generally thought was good) from the underlying seasonal prediction data (which they generally thought was poor). They did, however, think the prototype was 'very' easy to use and that seasonal wind prediction demonstrated much promise.

Introduction and aim

This report details the evaluation of a seasonal wind prediction prototype, named Project Ukko, produced as part of the bigger EU-funded project EUPORIAS. The prototype aims to support users, primarily wind farm owners and wind energy traders, in making decisions. Key decisions include when to plan routine maintenance and where to invest in new windfarms. The prototype consists of an interactive map-based interface and incorporates a range of different information visualizations. These visualizations provide data to support users make seasonal wind predictions – predictions about how much wind there is likely to be in a particular geographical area in a coming season (winter, spring, summer or autumn). While seasonal wind prediction is still in its infancy, there is much promise that this form of prediction can provide additional or alternative insights to decision-makers than the (more established and already widely-used) short-term wind forecasts, and than historical wind data - data about how much wind there was in a particular geographical area in the past. The aim of the research detailed in this report was to conduct a formative user evaluation of the prototype – an evaluation with potential users of the prototype that could inform the improvement of the design of the prototype.

Critical context

Importance of a user-focused approach

Vaughan and Dessai (2014) highlight there has been very little evaluation of climate service tools with potential users. While the predictability of the data used by climate service tools is often formally or informally assessed (e.g. by seeing how well predictions fare when compared to historical data or future observations), evaluation of the tools themselves is often neglected. This is a pity as potential users can often provide useful input to inform the design and improvement of the climate tools they use. Evaluating with users also allows developers of climate service tools to ascertain how successful their tools are – from the people that matter, potential users.

Climate service tools often incorporate complex data and data models. While developers strive to make these tools 'intuitive' and 'easy to use,' achieving this in practice is often not easy. Potential users, although usually domain experts, may not always understand all aspects of a complex information visualization interface and, in particular, the nuances of the way data is visually presented. A lack of understanding of what the interface is showing and how to interact with it does not signify ignorance on the part of the user – but that aspects of the interface are confusing or difficult to interpret. Our challenge as developers of complex tools is to make them as easy to interpret as possible.

Understanding users' understandings – A 'mental models' focus

The main focus of this research was on understanding potential users' understandings of the prototype interface – their 'mental models' (Johnson-Laird, 1983) of the prototype and how it works. While a mental models focus for user research has been adopted previously in Human-Computer Interaction research (e.g. Makri et al., 2007) and some structured methods exist to support this research (e.g. CASSM - see Blandford et al., 2008), this focus is relatively uncommon and has not, to our knowledge, been previously used to evaluate climate service tools with potential users.

Understanding users' understandings of complex interactive systems by attempting to elicit their mental models of those systems is an approach advocated by Norman (1983). Norman highlights that a user's 'mental model' of an interactive system evolves through their interaction with it, guiding their usage. Therefore an understanding of a user's mental model of a system can help explain the rationale behind how they interact with it (their interaction behaviour) and, potentially, their views on the success of the system. According to Norman, understanding users' mental models of interactive systems both has predictive and explanatory power. Norman had observed interactions with a broad variety of interactive systems and notes that *"most people's understanding of the devices they interact with is surprisingly meagre, imprecisely specified and full of inconsistencies, gaps, and idiosyncratic quirks"* (p.8). The more complex interactive systems are, the greater potential for users to misinterpret aspects of the functionality supported ('what the system does') and how to make effective use of that functionality ('how the system works').

Card and Moran (1986) suggest that having a detailed mental model of an interactive system can aid decision-making. While exposure to a system can help users create a mental model of how it works and how to use it, exposure does not guarantee the development of a detailed mental model. Indeed, Card and Moran (1986) found that even 'expert' users of interactive systems often had shallow models of how the system worked. This highlights the sizable challenge of ensuring understandability in complex systems such as the seasonal wind forecasting prototype that is the focus of our report.

Evaluating multiple 'layers' of interactive visualization tools – a 'nested model' focus

Another useful theoretical foundation for framing the user evaluation of interactive visualizations is Munzner's (2009) nested model for visualisation design and validation (see figure 1). When evaluating an interactive visualisation tool according to the model, the **domain problem characterisation** level involves us assessing how well the tool **supports important domain-specific user tasks**. The **data and operation abstraction** level involves us assessing whether the tool **supports the right types of data and operations** (generic rather than domain-specific tasks). The **visual encoding and interaction design** level involves us assessing the usefulness of how the data is presented visually and how well the interface supports users in interacting with the visual data. The **algorithm design** level involves us assessing the quality of the underlying algorithm that drives the

visualisation (and particularly whether it is fast enough to facilitate smooth, fluid interaction and timely updating of visual data).

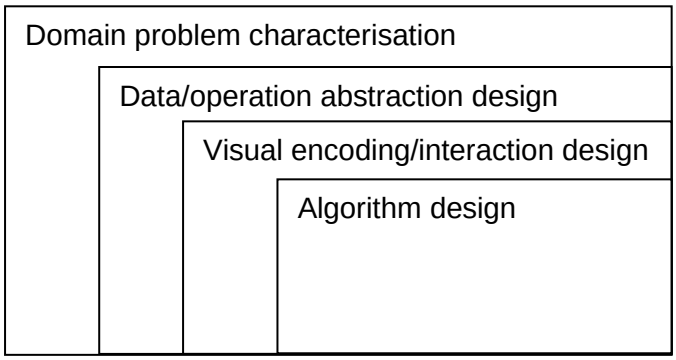

Figure 1: Nested model for visualisation design and validation (adapted from Munzner, 2009)

In the evaluation described in the report, we leverage Munzner's model by asking questions related to the various layers of the model during our evaluation. For example, to determine if the prototype supports important domain-specific user tasks we asked users 'what types of tasks would you like to perform when using a seasonal wind forecasting tool?' and noted whether and how the prototype supported the tasks they mentioned. To determine if the visual encoding and interaction design was sound, we asked users several questions about their understanding of the visualizations used at the interface and about the interface itself. To determine users thought the algorithm design was satisfactory, we asked users a question about the responsiveness of the prototype. See appendix A for a full list of representative questions asked during the evaluation.

An interesting challenge for evaluating interactive information visualization-based interfaces is that issues that manifest upstream in Munzner's model cascade downstream. For example, if a visualisation tool aims to support the wrong tasks (i.e. tasks that users do not want to perform), then even a sound interaction design and efficient algorithm is unlikely to result in users regarding the tool as successful. Also, Munzner (2009) notes that the nested levels in her model can overlap and this can make pinpointing the source of usability issues difficult; *"a poor visual encoding choice may case doubt when testing a legitimate abstraction choice, or poor algorithm design may cast doubt when testing an interaction technique"* (p. 922). This has important implications for our evaluation; issues with the prototype may have 'root causes' further upstream. It would therefore be useful to seek the root causes of some of the interface issues discussed in this report, time permitting, as this may result in the greatest overall holistic improvement of a prototype under evaluation. Another important implication of overlapping model layers is we cannot assume that addressing a particular interface issue will automatically result in the holistic improvement of a tool. It would be advisable to make design improvements, then evaluate the prototype again with different potential users to determine whether or not the improvements have resulted in a more successful tool. There is therefore scope for the EUPORIAS team to carry out a summative evaluation of the final version of the prototype.

Munzner's (2009) model highlights 4 potential threats to the validity of visualisation design that correspond to each layer of her model. As we mentioned earlier, we asked questions during our evaluation to understand the success of the prototype in relation to these layers (and therefore in relation to these validity threats, which are related to the layers). The threats are:

1) **Wrong problem** – users don't do that; the domain problem is mischaracterised. This emphasises the importance of interviewing and observing users from the target domain to verify the characterisation (something we did in this evaluation).
2) **Wrong abstraction** – designers are showing users the wrong thing; the chosen data types and operations do not solve the characterised problems of the target users. While it is common to assess this validity threat by asking target users to use a visualisation tool (which we did in this evaluation), it is also possible (and perhaps desirable) to do so through a longer-term field study and this is an opportunity for future work.
3) **Wrong visual encoding/interaction** – the way designers show it doesn't work; the chosen design is not effective at communicating the desired abstraction to the user. Munzner (2009) points out that 'expert reviews' undertaken by Human-Computer Interaction professionals are one way of assessing this threat to validity. We conducted an informal expert review of the visualisation prior to the user testing and recommended some changes to the interface and visual encoding as a result. Munzner also points out that it is possible to assess this validity threat by conducting user studies. While she describes the possibility of conducting laboratory-based studies with prescribed tasks, we opted to carry out a naturalistic user study outside of a lab setting - where participants carried out self-chosen tasks. This was because allowing them to choose their own domain tasks allowed us to evaluate at all levels of the nested model as part of a single evaluation study. Using prescribed tasks would not have allowed for testing at the domain problem characterisation level.
4) **Wrong algorithm** – the code is too slow; the algorithm is suboptimal in terms of time or memory performance. While Munzner suggests assessing this threat to validity by examining the computational complexity of the algorithm, we chose to evaluate the algorithm design by asking users questions about their perceptions of the responsiveness of the tool (allowing for an in-depth analysis of algorithm complexity if any issues were identified).

Evaluating from a traditional human-centred standpoint: Usefulness, usability and use

As well as evaluating the success of the seasonal wind forecasting prototype from a 'mental models' and 'nested model' perspective, we also opted to incorporate more traditional human-centred success criteria into our evaluation. Vaughan and Dessai (2014) highlight the potential utility of evaluating climate services in terms of **usefulness** (incorporating usefulness for supporting decision-making), **usability** and likeliness of **use**, stating that if climate data is not tailored appropriately to a target user domain *"it will not be useful to or usable by decision makers. As a result, it will not be used."* We adopt these three success criteria in this evaluation.

Evaluating the tool, not the data

It is also important to note at the outset that it is difficult to completely separate a climate service tool from its underlying prediction data and data models; without accurate prediction data a wind forecasting tool cannot make useful predictions and therefore no matter how well-designed the interface, the tool may not be perceived by users as successful. Without sound interface design, users may struggle to make predictions using the tool – again impacting their perceptions of success. The importance of climate services making accurate predictions is highlighted by Vaughan and Dessai (2014), who state that there may be negative outcomes for farmers, and potentially their communities, when forecasts are inaccurate. However when evaluating a tool based on *seasonal* forecast data, which is still an emerging area in the field of climate services, it is important to try to separate the tool from the data as far as is possible and practical. We must keep in mind that the reliability of

seasonal wind prediction data is outside the control of the developers of tools that aim to seasonal wind forecasting. Therefore although the reliability of the underlying data and predictability of the underlying data models may impact on users' perceptions of the tool, improving the prediction data and data models is likely to be outside of the scope of this work. That said, in order to ensure the long-term success of seasonal wind forecasting, data modellers and climate service tool developers will need to work hand-in-hand.

Method

Aim of evaluation

The primary aim of the evaluation was to provide formative feedback to improve future versions of the prototype. The secondary aim was to provide an indication of the success of the prototype in its current form. The success criteria we adopted were usefulness, usability and likeliness of use.

Choice of sample

For the evaluation, we conducted think-aloud observations with 5 potential users of the seasonal wind forecast prototype. The participants were identified and recruited through the EUPORIAS project network. All participants stated that either they, or their end users/customers, have an interest in using digital tools to make seasonal wind predictions. The participants were:

- P1 – R&D manager for a consultancy firm providing services to the wind industry.
- P2 – reseller of energy forecasts to wind farm developers, currently selling short-term forecasts.
- P3 - works in energy assessment department for a global wind energy developer and operator.
- P4 – manager at a major renewables company. Makes wind power production forecasts and related budget forecasts.
- P5 - works for energy resource department of a large utility company. This department analyses wind behaviour data, including observational and prediction data, to inform the design of future wind turbines.

Observation procedure

Before the evaluation, we explained the purpose of our evaluation to participants – to understand their understanding of the prototype. We told them that we would be asking them to use the prototype in the context of a real or realistic self-chosen task and that we would be asking them questions as they used it. We informed participants this would help determine the success of the prototype, inform its improvement and also inform the development of future wind prediction tools. We also told them that in order to get an accurate view as possible of their understanding of the prototype, we would not be demonstrating it beforehand. We gave participants the opportunity to ask questions before they started using the prototype and afterwards, but we informed them that we would not be able to answer questions about how the prototype works during the observation as we did not want to bias their understanding, interaction behaviour or opinions of the tool.

Sessions lasted around 45 minutes and were conducted remotely, as the participants were geographically dispersed around Europe and North America. Participants accessed the prototype using either the latest version of Firefox, Chrome or Safari and shared their screen

and audio with the researcher using a videoconferencing tool of their choice (typically Skype or Google Hangouts). We recorded both the screen and audio using Camtasia Studio. Participants gave their permission for the screen to be recorded and were made aware that they could:

- Stop the study at any time without penalty.
- Ask to review their recording or request that it is edited or deleted at any time.

None of the participants requested to withdraw from the study, review or delete their data. We asked participants if they had previously used the prototype or earlier versions. Only P4 had used an earlier version, which was only 'for a few minutes.' There was nothing in P4's observation data to indicate that he had been at an advantage, or disadvantage (in terms of forming a complete and accurate mental model of the prototype) from having used the prototype briefly.

We began the evaluation by asking participants questions about their context of use for the potential prototype. We asked them to tell us their job role and how a seasonal wind forecasting prototype might assist them in their jobs. As none of the organisations the participants worked for currently use seasonal wind forecasts, we asked them why not. We also asked about the different types of tasks participants would like to perform using a seasonal wind forecasting tool and the different types of information they would want it to provide them.

Next, we asked them to explore the prototype in the context of a realistic task. We stated this might be, but did not necessarily have to be, one of the tasks they identified previously. We asked participants to 'think aloud' while they were interacting with the prototype – telling the researcher exactly what they were doing and why. We asked questions opportunistically to probe participants' understandings of the prototype. Our 'opportunistic' questions were adapted on-the-fly from a script we prepared (see appendix 1). These were questions about what different aspects of the prototype meant or what their function was. We focused in particular on the key functionality of the prototype – the different icons on the legend (skill, predicted strength, predicted change, installed wind power), the observations and predictions graphs and the 'skill' and 'installed wind power' bars on the pop-up console. Other questions probed the usefulness of the tips that pop up for explaining elements of the visualization (tooltips), some of the options on the 'settings' menu and the responsiveness of the tool. These questions covered each of Munzner (2009's 'nested layers.'

After the observation, we asked participants some final questions focused on their perception of how well the prototype met the success criteria of usefulness, usability and likeliness of use; we asked them how useful and easy to use they thought the prototype was and how we could make the prototype more useful/usable. We also asked them how likely they were to use the prototype in the future and how we might make it more likely that they do.

We transcribed the audio and analysed it using a qualitative thematic coding approach – where we labelled the important themes from the data (e.g. potential for seasonal wind forecasts, reasons for not using seasonal wind forecasts so far etc.). These themes are used to structure our discussion of the findings in the next section. Many of the themes were directly related to the probing questions we asked. For example, we discuss findings related to the 'skill,' 'predicted strength' and 'predicted change' legends and to the 'observations' and 'predictions' graphs. We now discuss each of these themes and make design recommendations related to our findings. For each theme, we summarise the important findings at the beginning of each sub-section. Then we discuss the findings in detail, using direct quotations from the participant observations as evidence.

Findings and design recommendations

The findings and recommendations for improving the design of the prototype are presented in this section. In each sub-section, we include an 'in a nutshell' summary at the beginning which serves to summarise the most important findings. We begin by discussing participants' opinions of the **potential for seasonal wind forecasting** and their **reasons for not using seasonal wind forecasts so far**. These discussions set the context for our design-related findings. Next, we list the **real-world information tasks and information needs** that participants would like a seasonal wind forecasting prototype to support. This provides an insight into potential user requirements for seasonal wind forecasting tools; what they want to be able to do using these tools. Next, we discuss participants' understandings of **the skill, predicted strength, predicted change and installed wind power legends**. We focus in particular on where participants misinterpreted these legends and suggest ways of improving the design to reduce potential for these sorts of misinterpretations. We then discuss participants' understandings of the **observations and predictions graphs** and of **the skill and installed wind power bars** (which are on the same pop-up console as the graphs). Next, we discuss participants' opinions of the **informational 'tooltips'** on the console. Then, moving away from the console and legend, we briefly discuss issues and general design recommendations related to the **settings menu**, **zoom functionality** and **responsiveness** of the tool. We then discuss participants' comments on **additional functionality** they would like to see added to future versions of the prototype. Finally, we discuss **how successful** participants perceived the current version of the prototype to be in terms of Vaughan and Dessai's (2014) user evaluation criteria of **usefulness**, **usability** and **likeliness of use**.

Potential for seasonal wind forecasting

In a nutshell: Seasonal wind forecasting has much future potential.

- P2 commented that he thought the EUPORIAS project is a 'pioneer' in respect to providing seasonal rather than shorter-term forecasts and stated that seasonal forecasts with high predictability 'would come' but are 'not here yet.'
- P3 highlighted the great potential for seasonal wind forecasting, citing the impact of seasonal forecasts in North America in 2015. He stated there has been an intense 'El Nino' effect this year which has seen a warming of the Central Pacific ocean's surface. P3 stated that it has been hypothesised that the El Nino effect impacts on wind speeds. This suggests that changes in wind conditions in recent years may make it particularly important to make seasonal (as opposed to short or long-term wind predictions).

Reasons for not using seasonal wind forecasts so far

In a nutshell: Seasonal wind forecasts are not currently used because of the current lack of predictability and reliability of the underlying prediction models.

- The majority of the participants were not yet convinced of the value of seasonal forecasts due to the lack of predictability and reliability of the underlying prediction models. For example, P4 stated that the utility of seasonal wind forecasts depended *"on the uncertainty you give to me"* (P4).
- P1 stated that seasonal wind forecasts had not yet proven their usefulness *"due to lack of reliability"* (P1) of seasonal prediction data. *"It could be that right now it's not any more accurate for making decisions than just flipping a coin"* (P1). *"If we could through a case study that there is a way to evaluate with a reasonable degree of*

accuracy what the production is going to be for the next 3 months, then I think this could be valuable information…" (P1).

- Similarly, P2 commented that his energy prediction resale company did not currently re-sell seasonal wind forecast data because *"seasonal forecasting has not demonstrated its predictability yet"* (P2). *"For short-term forecasting, predictability is sometimes poor. So imagine how it is for seasonal forecasting!"* (P2).
- P3 commented that seasonal wind forecasting is *"a relatively new technology with lots of uncertainty – much more than short-term forecasts. So there is still scepticism…"* (P3). P3 thought that this scepticism existed because it was more difficult to make seasonal versus short-term wind forecasts. Similarly, P5 noted that use of seasonal wind forecasts had not become commonplace yet in the energy industry because they are *"quite difficult to calculate"* (P5). "*Meteorological forecasts are usually useful only for 10 days. So trying to do that on a seasonal basis is quite complicated"* (P5).

Possible real-world usage (information tasks and needs)

In a nutshell: Seasonal wind forecasting tools can be used to make decisions related to energy trading, wind farm investment and maintenance. They can also be used to predict wind farm performance and profitability.

- P1 stated that seasonal forecast data might be used to **predict wind farm performance and profitability**. He stated that a representative task might be: *"I know a wind farm is going to be built in the coming month in this region. What could I expect in the first month of production?"* (P1).
- P1 mentioned that his clients may use seasonal wind forecasts to support initial prospections – where the seasonal forecasts can be used to **identify potential geographical regions for development** (e.g. when a developer wants to build new wind farm projects in a specific area). Seasonal forecast data might be used by developers in order to decide whether and where to prospect (and potentially build). He stated the prototype could be used, in particular, when undertaking energy yield assessments for the purpose of preparing a case for negotiating loans with banks for new wind farm projects.
- P1 also mentioned that the prototype might be used to make **energy trading decisions**. He stated: *"if a particular region is likely to produce high energy this winter, maybe this will make you more likely to buy energy from this region"* (P1). *"This next winter will be 20% more windy and therefore x% more productive than a normal winter." I think we're really after this kind of information* (P1).
- P2 envisaged his resale customers using seasonal forecast data for making **maintenance decisions**: *"Maintenance people are rough people; they are not scared of bad weather. But if you stop a wind turbine for maintenance and it is during windy days, then a lot of money will be lost… So maintenance teams try to plan maintenance operations during low wind periods. So if you could forecasts that the next spring will be less windy than usual, then maybe it's a good idea to planify a lot of maintenance operations for that time"* (P2).
- P5 also mentioned the potential for seasonal forecast data to be used for making maintenance decisions. He stated they might be able to predict windstreams or extreme weather conditions and pre-**plan purchases of the necessary maintenance parts** to mitigate typical wind farm breakdown issues: *"if you can understand that a season is going to be probabilistically windier than others, maybe there will be some extreme wind conditions in that season. So you can be aware in advance of what might happen. So you might purchase spare parts for the wind farms in case of accidents"* (P5).

- P2 also envisaged resale customers using seasonal forecast data for making **investment decisions**. He commented that this data could be useful for financial planning: *"We are not wind farm owners, but I could imagine being in their shoes. If we could anticipate a good income for the next season, this could help us to manage our assets and say 'we will not need this loan for the next month' or 'we can wait until the wind is coming to borrow or invest money.' You could make your investors happy by saying 'don't worry, next month there will be a lot of money in your pockets because there's a lot of wind forecast"* (P2).
- Similarly P4 stated that seasonal wind predictions might help make decisions related to **wind farm construction**. Knowing there was likely to be an unseasonably windy or not-so-windy spring, for example, might lead to energy companies such as his deciding to start their construction work sooner rather than later (or vice versa). P4 explained that *"you can't just construct a wind farm any time. You need a minimum condition of wind, minimum condition of rain"* (P4). Seasonal wind forecasts might result in better estimates of when construction conditions are likely to be good which, in turn, could lead to cost savings (as construction work will not be planned for unsuitable days/weeks).
- P3 also thought seasonal forecasts could be useful for making investment decisions and, particularly for **assessing financial risks**. He noted that an awareness of unusual seasonal wind conditions could alert assessors to where wind energy production (and therefore income and profitability) may differ from their financial plans. It could also assist the writing of **financial plans** in the first place. Similarly P5 thought that seasonal wind forecasts could provide a prediction of wind variability in a geographical area, which could be used to make financial plans.
- Also related to making financial decisions, P5 thought seasonal wind forecasts would be useful for making **cashflow predictions**: *"it would be good to be able to check if a wind farm is going to produce an excess of energy in advance so you can update your future cashflows and know how much money you have in advance"* (P5).
- P4 stated that seasonal wind forecasts could provide an indication of **potential future income**. He stated that seasonal forecasting has the potential to improve their income predictions as, at the moment, these predictions are made based on averages rather than on sophisticated predictions.

Skill legend

In a nutshell: Participants generally understood the skill legend, although some voiced uncertainty about what it denoted. Some participants struggled with understanding why negative skill levels were shown sometimes but not others.

One participant noted that looking at the skill legend was particularly useful, particularly when looking for where there are both existing installed wind farms *and* the prediction data for that area has a positive skill level:

- *"What strikes me in the interface is that usually where there are turbines there is no skill."* (P1).
- The participant proceeded to find an example of where there was a wind farm installed and a positive skill level, mentioning that these were often areas of interest for post-construction and grade-integration (see figure 2).
- P1 commented that *"for this, you need to have at least slightly positive skill where there are wind farms"*. He suggested this might be an interesting area for people who own or operate wind farms because *"if currently the predictions for the next few weeks are quite low and you expect to have a high winter, maybe you will move towards some preventive or regular maintenance that will involve cutting the turbine*

down. If the meteorological forecast for the next 2 weeks is quite low, you expect the coming season to be quite windy and you have upcoming maintenance, you might want to do it in the coming weeks because that's when you'll lose less money turning the turbine off." (P1).

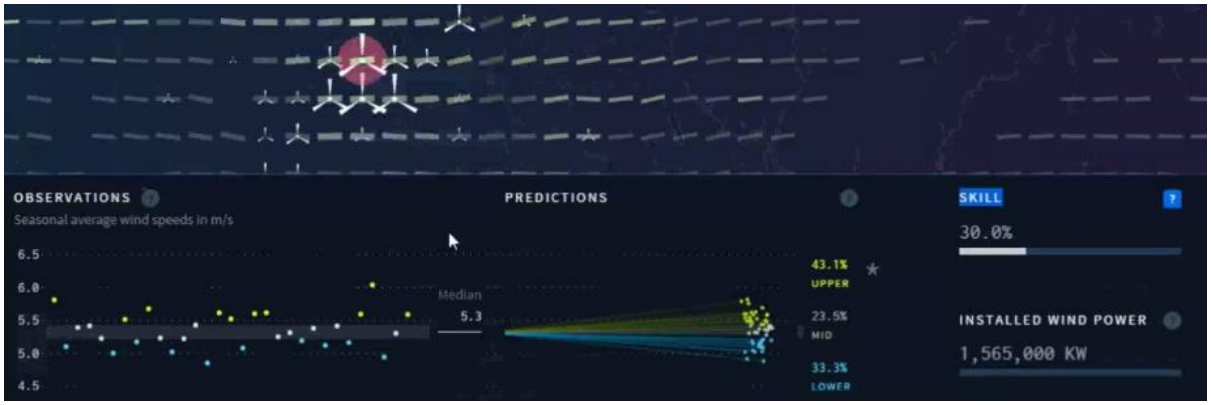

Figure 2: The co-located wind farm installation and prediction selected by P1

Most participants interpreted the skill legend correctly, although their explanations of it varied and both P2 and P3 voiced some uncertainty:

- P4 stated that the skill legend was *"useful because you can see the regions around the world where you have better predictions"* (P4).
- P1 thought the skill legend indicated *"where your model performed well."* Similarly P3 commented that a high skill level meant he could *"trust the forecast more than in another area."*
- P2 commented *"skill is the predictability, isn't it?" How well the predictions compare to reality"* (P2).
- P3 commented that *"what I would really like to know is what skill really means. So far, I've interpreted it to mean how much I can trust this forecast"* (P3).
- **Design recommendation 1 (DR1): Consider including informational 'tooltips' for legend items as well as for items in the popup console window.**

Participants generally understood that if there was no rectangular icon on the map screen in a particular location, this meant the skill level in that region was too low for the prototype to display prediction data related to that region:

- *"I can see skill somewhere [clicks on various icons on the map], but not everywhere. So I guess where it's negative it doesn't appear at all"* (P1).
- Similarly P3 stated he thought the lack of prediction data in certain areas on the map was because *"the skill is so low that the model cannot provide a prediction"* (P3).

However, there was some confusion about whether rectangular icons with a negative skill value were displayed on the interface or not:

- When asked why he had previously seen a skill level of -5% displayed on the console, P3 stated this was because only those rectangles with a skill level of zero were not displayed; negative skill levels were. This is a misinterpretation caused by the inconsistency of the tool in displaying prediction data where wind farms currently exist, even if the skill level for that data is negative and not displaying prediction data where no wind farms currently exist if the skill level is zero or less. **DR2: Consider whether or not to display prediction data where wind farms currently exist and the skill level is negative.**

One participant also appeared to be confused by the use of both colour and opacity on the legend:

- P5 said he was unsure whether the skill legend referred to the rectangles on the map, or to the colour of all the different icons on the map. He was also unsure what the rectangles on the map were for in general. **DR3: Consider using different rectangle colours to denote difference in skill level, different rectangle thicknesses to denote difference in predicted strength and different sloping angles to denote predicted change.**

Predicted strength legend

In a nutshell: Participants were confused by the concept of 'predicted strength' on the legend bar. This could be renamed to 'predicted wind speed.'

Most participants were confused by the concept of 'predicted strength' on the legend bar:

- P1 commented *"I don't know what strength will mean*." When asked what he thought it *might* mean, he paused for several seconds and stated *"strength would be the mean wind speed compared to the global mean of predictions?"*
- P2 misinterpreted the 'predicted strength' legend, confusing it with predicted change. He said: *"predicted strength. This is something I don't understand. Is this the amplitude of the change? If the weather will deviate more or less in respect of the climate averages. Is that right? The wider the rectangle, the more the prediction deviates from the climate"* (P2). He clarified that this meant that the weather will be different from what it was in the past.
- P3 was initially 'not sure' of the difference between the 'predicted strength' and 'predicted change' legends, but assumed that predicted strength meant *"how the wind speed will be next month, for example"* and predicted change *"how wind will change in the long term"* (P3). Even after reading all the tooltips on the console, he commented that he was still unsure of the difference.
- **DR4: Consider renaming 'predicted strength' to 'predicted wind speed.' This would need to be verified with several potential users.**
- P4 mentioned he could 'not distinguish' between many of the rectangular icons on the map in terms of predicted strength and spent a few seconds scrolling around the map to find some rectangles that were much thicker than others. Once he did, he compared the data on the console for both and could not discern a difference (possibly because he did not hold his mouse cursor over any of the individual predictions which would tell him the predicted wind speed for each of the 51 simulations). **DR5: Consider changing the bandings for different rectangle thicknesses for 'predicted strength,' to display a greater variety of rectangle thicknesses on the map. DR14: Consider highlighting the relevant legend item icon to indicate how the selected map icon relates to the legend (e.g. if a prediction with the thinnest 'predicted strength' rectangle is selected on the map, the thinnest rectangle is selected on the 'predicted strength' legend item).**
- It became apparent that the predicted strength legend was not clear to P4 because, when comparing a prediction with a thick rectangle (denoting 'high predicted strength') with that of a thinner rectangle (denoting 'low predicted strength'), he tried to compare the prediction percentages for upper, mid and lower terciles for both rectangles rather than the individual predictions. A few seconds later, however, he did compare the range of the individual predictions for both rectangles with each other to understand the 'predicted strength' legend. However, this did not help him to successfully revise his interpretation; he commented that *"the thicker the rectangle the more dispersion in the observations and predictions" (*P4).

- P5 was 'not really sure' what the predicted strength legend icons meant.

One participant suggested renaming 'predicted wind strength' to 'predicted wind speed,' as 'speed' was more commonly used in the industry:

- P1 suggested renaming 'strength' to 'predicted wind speed' because 'strength' is *"not commonly used in the industry"* (P1). P3 made a similar comment and thought 'predicted wind speed' would be clearer. **DR4: Consider renaming 'predicted strength' to 'predicted wind speed.' This would need to be verified with several potential users.**

Another participant suggested using colours instead of rectangle thickness to denote predicted wind strength:

- P2 suggested that, for the predicted strength legend, using colours instead of rectangle thickness might be useful – with red to denote increases and blue to denote decreases. He stated that this was 'standard in the industry.' He was aware that it would be necessary to model 3 different variables on the map though and so understood that this change may not be possible or desirable.
- We do not list this as a design recommendation as we later make a recommendation that the *skill* legend might make use of colour instead of the 'predicted change' legend to allow the three legend items to be better distinguished from one another.

Predicted change legend

In a nutshell: Most participants understood the 'predicted change' legend item, but there was some ambiguity in some of their understandings.

The 'predicted change' legend item was understood better by participants than the 'predicted strength' legend item:

- P1 demonstrated his understanding of the 'predicted change' legend by selecting a blue, backward-sloping rectangle, and stating that this was likely to mean that the prediction for a future winter was much lower than the mean of previous observations and a green forward-sloping rectangle, stating that the prediction for a future winter was much higher than previous averages.
- Similarly P4 commented that 'predicted change' was whether in the future there would be more or less wind than the average of the previous observations and P5 stated that it was whether the predicted values were higher or lower than the historical average.

But there was still some ambiguity in some of their understandings:

- P3 was initially unsure of what the 'predicted change' legend was showing, but commented that he thought it was probably *"how far the predicted wind speed for next month will be from the long-term median of the history"* (P3). **DR13: Consider including a caption on the legend to explain 'predicted change.'**
- This participant frequently used the word 'month' instead of 'season,' suggesting the potential to make it clearer at the interface (a) that predictions are for a particular season and that (b) seasons are spring, summer, autumn and winter.
- This suggestion is also supported by evidence from P4, who asked *"what do you mean by seasonal?"* The notion of 'seasonal' is inherently ambiguous; it may be useful to explain on the interface (e.g. in the 'predictions graph' tooltip) which months class as spring/summer/autumn/winter. **DR6: Consider explaining on the interface what 'seasonal' means and which months correspond to each season.**

- P2 suggested changing the visual representation of predicted change so that the rectangles no longer sloped. He commented that the existing visualisation *"is original and it's elegant, but it's not within the standards of the sector"* (P2). He suggested using colour transparency/opacity to model predicted change, with different colours to model predicted strength. He did, however, confuse the meanings of the 'predicted strength' and 'predicted change' legend items, so this suggestion would require adaptation if implemented to take this into account. His confusion is evident in the following comment: *"these two variables [points cursor at 'predicted strength' and 'predicted change' legend items] are actually only one. They are the same thing. If it increases a lot, then I think the predicted strength is high."* (P2).

Some participants made specific suggestions for improving the 'predicted change' legend item:

- *"I myself would call this 'next winter trend' or something like that, but perhaps many people would not agree"* (P1). **DR7: Consider whether or not it is appropriate to rename 'predicted change' legend item to 'next winder trend' or similar. This would need to be verified with several potential users.**
- P4 suggested the potential utility of explaining the legend items on the interface, perhaps by including a similar question mark icon beside each legend item to that used to denote the availability of tooltips on the console. Similarly, P5 stated *"I would like to know why there are different symbols in different areas, and what they mean. This is the information I'm requesting"* (P5). **DR1: Consider including informational 'tooltips' for legend items as well as for items in the popup console window.**

Installed wind power legend

In a nutshell: Most participants understood the 'installed wind power' legend item itself, but were confused with the ways the icons were used on the map.

The meaning of the 'installed wind power' legend item itself did not confuse participants and the icon was generally regarded as straightforward to understand:

- Although he struggled to interpret the skill, predicted strength and predicted change legend items, P5 found the *'installed wind power' legend easy to understand: "the bigger the icon, the bigger the wind power installed"* (P5).
- P1 suggested this legend made sense to him, and thought it was 'perfectly understandable,' but he thought *"wind capacity is certainly more precise than wind power"* (P1). **DR8: Consider renaming 'installed wind power' to 'installed wind capacity' or something similar. We suggest 'wind power installed in region.' This would need to be verified with several potential users.**

But the usage of the 'installed wind power' icon on the map did cause confusion amongst some of the participants:

- When he clicked on a part of the map with both a wind prediction and installed wind power, P3 was initially confused: *"it looks like there are some windmills here, what's this?"* (P3). However, after some exploration, this became clearer for him: *"I guess the only difference is that the rectangles without windmills have no installed wind power"* (P3).
- P3 was also curious about whether the installed wind power icons displayed in the map were all the wind farms installed in the world, or just a sample. Similarly P4 initially referred to the wind farm icons as *"a wind farm"* rather than as a group of wind farms. However, after reading the 'click to see details about this region' tooltip, he revised his interpretation and stated that the purple shaded circle denoted a

region and that the observations and predictions data was likely to refer to all wind farms in the area bounded by the circle . These comments, combined with P2's comment about also wanting to see forecast installations on the map, indicate a potential misinterpretation that the icons represent individual wind farms rather than provide an indication of installed wind capacity in a specific geographical region. **DR8: Consider renaming 'installed wind power' to 'wind power installed in region.' This would need to be verified with several potential users. DR9: Explain in an informational tooltip for 'installed wind power' that the icons represent wind capacity in a specific geographical region, *not* individual wind farms.**

- Although the installed wind power legend item shows 3 different sizes of icon (low, medium and high), the icons on the map are currently generated at more than 3 relative sizes (see figure 3) below, which illustrates several sizes of windmill icon. While none of the participants mentioned this, it could lead to interpretation issues and to issues in spotting installed wind power; figure 3 shows what is displayed when a user zooms as far as possible into eastern Romania. The selected installed wind power icon (denoted by the circle) is extremely small and therefore may be difficult for users to notice. **DR10: Consider using only 3 different sizes of 'installed wind power' icon on the map.**

Figure 3: Installed wind power icons of various sizes may be difficult to interpret and smaller ones difficult to spot

- P5 asked *"why do I need predictions in areas where there are no wind farms installed or if there are no future prospects of having wind power installed...it's not useful for us, but maybe it's useful for other people"* (P5). **DR11: Consider whether it is necessary to include predictions in areas where there are no existing or prospective wind farms installed. This would need to be verified with several potential users.**
- P5 also wondered why there were predictions in some ocean areas but not others as he found this 'a bit confusing.' He asked *"is it because there is some problem with the predictions? Maybe you don't have enough to do a prediction in some places and in those places you don't show anything.* (P5). When asked to clarify whether he could see predictions in areas where he knew it would not ever be possible to create wind farms, P5 answered 'yes.'
- P2 commented that it was 'good' to show installed wind power icons on the map. But he would also have liked to have seen forecast installations on the map (i.e. not only wind farms that have already been installed but also those currently under construction). **DR12: Consider whether or not to show individual existing wind**

farm installations and possibly forecast installations on the map. This requires a decision to be made on the relative benefits and drawbacks of switching from the icons representing installed wind capacity in a particular geographical area to representing individual wind farms.

Suggested new legend

In a nutshell: We present an illustrative suggested new legend incorporating many of our legend-related design recommendations.

A potential legend incorporating several of the design suggestions related to each of the legend items is presented in figure 4. Note that this re-design is intended to be illustrative rather than prescriptive and there is scope to change many of the fine details (e.g. the colours used, means of highlighting the legend items etc.). In figure 4, several of the legend labels have been renamed to address DR4, DR7 and DR8 and a caption explaining 'predicted change' has been included to address DR13. The skill level is no longer denoted by opacity, but by colour – using a 'traffic light' colour system to denote percentage skill. Tooltip question mark icons have been included to explain each legend item further (to address DR1). Scales have been included for each legend item to help users understand the concrete relative differences behind 'low' and 'high,' 'increase' and 'decrease' (addressing D22, discussed later). Boxes have been drawn around the applicable legend icon for the selected map icon to make it clearer where on the scale the selected icon on the map relates (to address DR14). These boxes could potentially be replaced with purple-shaded circles or rectangles to ensure consistency with the shaded circles that appear when an icon has been selected on the map.

Figure 4: Suggested new legend based on participant misinterpretations and feedback.

Observations graph

> ***In a nutshell: The observations graph was fairly well-understood by participants, but several ambiguities came to light when we probed their understanding of the data presented.***

The 'observations' graph on the pop-up console that appears when a user clicks on an icon on the map was fairly well-understood by participants. One participant immediately related to the multiple predictions provided as an 'ensemble,' finding the graph easy to interpret because it followed industry conventions he was already familiar with:

- P3 commented that providing multiple predictions (an 'ensemble'), where *"each ensemble member provides a different forecast"* (P3) was similar to what is usually done when making short-term wind predictions. Therefore it is likely this follows industry conventions. P5 commented that the ensemble was 'quite nice' as the dots showed where the predictions lay in relation to the median.

However, several ambiguities came to light when we probed participants' understandings of the data presented on the observations graph:

- P4 was unsure whether the prototype was using observation data from the same seasons in previous years (e.g. winter) to make a prediction for winter 2012, or whether observations data from other seasons (i.e. spring, summer or autumn) was also used. **DR15: Consider explaining in the observations graph tooltip whether the graph uses observations data from successive same seasons to come up with predictions for that same season in the coming year (e.g. uses winter observations to make a prediction for next winter), or whether it uses averages from across seasons).**
- P1 assumed that the observations graph was displaying average wind speed observations from successive winters from 1981-2011 (in order to forecast a predicted wind speed for winter 2012), rather than displaying average wind speed observations across all seasons in 1981-2011. This aspect of the interface is potentially ambiguous. He commented: *"everywhere you refer to 'seasonal,' but it's only here [highlights 'Winter 2012' under 'predictions' graph] where you can find out that seasonal actually means winter. And I could not find out if I could change the season anyhow."* He suggested the need for users to be able to change the season the forecast would be predicted for. **DR16: Consider giving users the option to change the season for which the predictions are for (currently always 'Winter 2012').**
- Several participants commented that they would expect the prototype to make forecasts for a coming season (currently the forecast is for a season in the past (Winter 2012) rather than a future season (e.g. Spring 2016). **DR16: Consider giving users the option to change the season for which the predictions are for (currently always 'Winter 2012.'**
- P2 commented that *"points are more elegant than lines, but the industry is more accustomed to lines*." **DR17: Consider a line graph view instead of or in addition to a scatterplot view. This would need to be verified with several potential participants.**
- P3 stated he didn't know whether the observations were based on recorded local observations or from the analysis dataset – a public database that includes wind speed. When he viewed the 'installed wind power' tooltip towards the end of his session, the Windpower.net acknowledgement confirmed that the analysis dataset was used.
- P1 suggested being able to alter the level of granularity of the forecast period: *"Seasonal means spring, summer, autumn or winter. But months could be interesting*

to introduce, and potentially years." (P1). We do not include this as a design recommendation as we are aware that the scope of the prototype is intentionally restricted to *seasonal* (as opposed to monthly or annual) wind predictions.

Predictions graph

In a nutshell: The predictions graph was also fairly well-understood by participants but, as with the observations graph, several ambiguities came to light when we probed their understanding of the data presented.

Like the observations graph, the predictions graph was also fairly well-understood by participants. For example:

- P4 correctly interpreted the predictions graph and was able to come up with an interpretation of the predictions (in this case 23.5% were in the upper, 39.2% in the mid and 37.3% in the lower tercile). He commented that this meant that it was likely that winter 2012 would have either average or below-average wind.
- Similarly, P1 correctly interpreted the predictions graph shown in figure 5 (which showed an upper tercile of 21.6%, mid tercile of 31.4%, lower tercile of 47.1% and skill level of -4%).

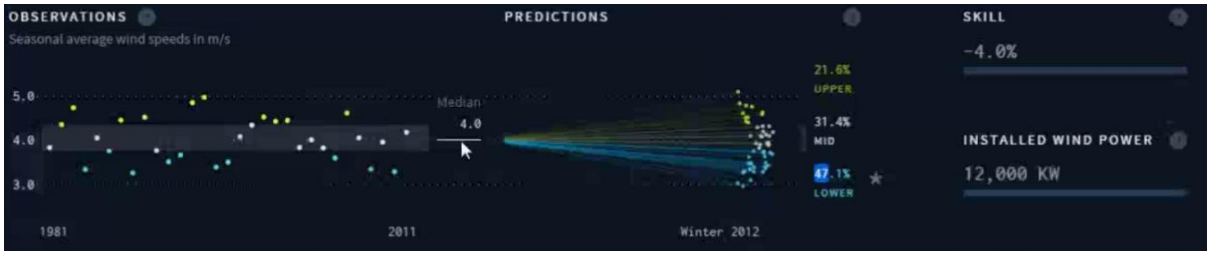

Figure 5: The wind farm installation selected by P1

- However, he voiced some ambiguity about what the intervals displayed in the graph meant: *"I can tell from the graph that there's a 21% chance of being in the upper interval, whatever the interval is."* (P1). **DR18: Consider mentioning that the intervals are terciles in the 'predictions' informational tooltip.**
- Some additional ambiguity arose when he explained the predictions graph further: "*there is a 31% chance to be within the standard deviation, I guess and 47% to be lower than the mean minus the standard deviation."* (P1). However, the participant did make a reasoned decision based on the predictions graph (see figure 5); he noted that in the first month, he would have a lower payback than average next year if investing in creating a wind farm in a particular area. He mentioned that this would not put him off suggesting that his clients invest in a particular region, but if external constraints were likely to cause delays of several months, he would advise customers that now would not be a particularly good season to build.
- P1 asked whether winter 2012 in figure 5 would be lower than the median of all observations from 1981-2011 or the median of observations taken during *winter* of 1981-2011. **DR15: Consider explaining in the observations graph tooltip whether the graph uses observations data from successive same seasons to come up with predictions for that same season in the coming year (e.g. uses winter observations to make a prediction for next winter), or whether it uses averages from across seasons.**
- P1 requested the ability to be able to select a season (e.g. 'Spring 2016') for the tool to make predictions for: *"a breakdown where I can select the period I want to look at"*

(P1). **DR16: Consider giving users the option to change the season for which the predictions are for (currently always 'Winter 2012.'**

- P2 thought the predictions graph was *"very clear."* However, he referred to the median figure as 'the mean' and then asked *"why is this the median and not the mean? I think people will not understand what the difference is between the median and the mean. If you show the median instead of the mean for some reason, this should be explained…"* (P2). **DR19: Consider either using mean instead of median to denote average in the predictions graph, or explain in the 'predictions' tooltip why median instead of mean is used.**
- P5 referred to the median as *"the median of the prediction"* rather than the median of the *observation* data, suggesting possible confusion. **DR20: Consider explaining in the 'observations' tooltip that the median is calculated of all 51 observations on the observations graph.**
- P2 also stated that he was missing from the prediction graph "*the result of this ensemble [points cursor at upper, mid and lower tercile figures]"* (P2). He continued *"what is the result of combining everything? I need a big overall number here that says the prediction will be 20% more or less wind. This is too diffuse for me as a hypothetical decision-maker. A decision-maker needs a number. Our customers don't like uncertainties. They like simple numbers… In general people don't like uncertainties. They like to go to church and listen that if people are good they will go to heaven and if they are bad they will go to hell"(*P2). **DR21: Consider providing an overall interpretation of the prediction data – e.g. an overall prediction score and some on-screen guidance (e.g. in a tooltip) about how to interpret it.**
- Providing an interpretation of the prediction data might support P2 in his decision-making. It also might clear up some of the initial confusion P3 experienced; P3 clicked on an icon where 27.5% of predictions were in the upper tercile, 21.6% in the mid and 51% in the lower (see figure 6). He commented: *"here it says that the average wind speed is going to be 21%... let me check… around the historical median. But I don't know. I would need some explanation from the model"* (P3). A few seconds later, the data seemed to become clearer for P3, who commented that *"probably what I can interpret from here is that 51% of the models are saying that the wind speed will be lower than the long-term"* (P3). **DR21: Consider providing an overall interpretation of the prediction data – e.g. an overall prediction score and some on-screen guidance (e.g. in a tooltip) about how to interpret it.**

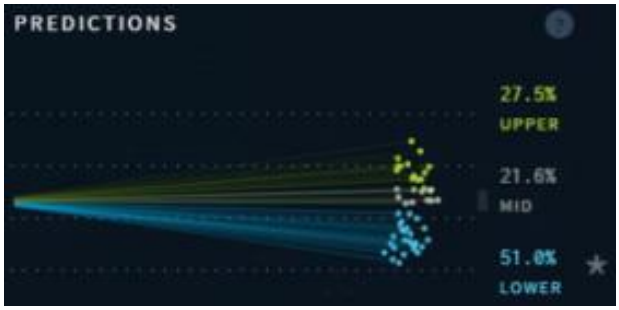

Figure 6: The predictions graph P3 looked at. He was initially confused about how to interpret the data

- Providing an interpretation of the predictions data might also help P5, who commented that although he understood that the grey star on the 'predictions' graph denoted the breakdown of the 51 predictions across the three terciles, he did not *"know what to do with this in this case"* (P5). In this case, the breakdown was 25.5% upper, 31.4% mid and 43.1% lower tercile). He commented that while the grey star (which appeared next to the 43.1% figure) told him that *"the prediction says that the wind for the next season will be lower than the climate average with a certain*

probability" (P5), *"it is not really easy for me to understand what to do with this right now"* (P5). P5 stated he thought the star *"was just pointing out the greatest score of the predictions I guess, but I'm not sure"* (P5). **DR21: Consider providing an overall interpretation of the prediction data – e.g. an overall prediction score and some on-screen guidance (e.g. in a tooltip) about how to interpret it.**

Skill bar

In a nutshell: The skill bar was well-understood by some participants and not others. Participants generally found skill levels difficult to interpret and make decisions with.

The skill bar (and associated skill percentage level) was understood by some participants, but not others. When understood well, several participants commented that the low skill level it depicted meant that they should not rely on the prediction for this geographical area:

- When asked what he thought the skill percentage meant, P2 replied 'good question!' He thought it meant that *"in the past, only 2% of the predictions were right. This is useless because it means that 98% were wrong."* (P2).
- After reading the 'skill' tooltip and when faced with a skill level of -4.0% for a selected wind farm, P1 commented *"this tells me that I should not rely on this prediction… because it's negative! It means that the model will perform worse than if I just assume it's a normal winter."* (P1). Similarly, after reading the tooltip, P4 commented that the skill level was *"a measure of uncertainty"* and that a negative skill percentage would mean that the prediction models would be worse than 'just making a guess based on historical data' (as described in the 'skill' tooltip).

When not-so-well understood, the skill bar proved to be a source of confusion amongst participants:

- P3 initially had 'no idea' about what the 'skill' bar was showing and did not notice the question mark icon denoting a tooltip. When asked what his guess would be, he stated *"it must be to do with predictability… how easy it is to predict… using this model"* (P3).
- When asked what a skill score of 11% meant in terms of 'how easy it is to predict,' P3 answered *"I don't know, I have no idea."*
- P3 did remember, however, that in a meeting he had learnt some heuristics about how to judge the boundaries. The prototype already includes heuristics to help users judge the skill level of a particular prediction (in the form of the skill legend). But it may be useful to make the boundaries of each skill level explicit on the interface (e.g. 'low' = 0-10%). The same could also be implemented for the predicted strength, predicted change and installed wind capacity legends (see figure 7). **DR22: Consider providing a scale for each legend item.**

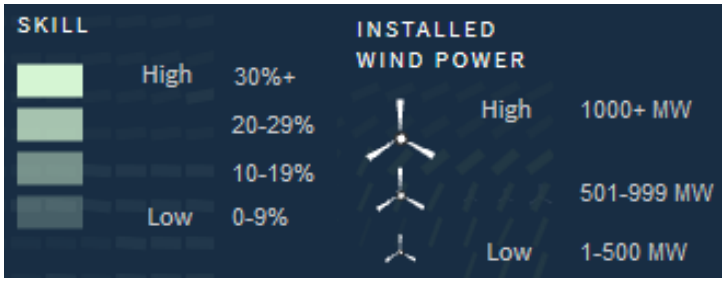

Figure 7: Illustrative example of making boundaries of each skill level and installed wind power level explicit at the interface. Note that these boundaries are approximations and may not be those used when developing the prototype.

- Similarly P4 commented *"I don't know how much you can achieve with this skill model. I don't have this information"* (P4). He went on to explain that it was not simply knowing the skill boundaries that would be important, but knowing the *expected performance* of the underlying prediction model: *"knowing, for example that it is not possible to achieve 100%"* (P4). **DR22: Consider providing a scale for each legend item.**
- P2 commented: *"for me, there is no intuitive threshold of when I can rely on this [points cursor to 'skill level' bar]… I have no idea if 34% skill level is useful or not. This is a very non-intuitive parameter. [Reads 'skill' tooltip]…I cannot translate this into what to do"* (P2). **DR22: Consider providing a scale for each legend item.**
- P2 stated it was 'a pity' that there was no data in a particular area of Central America around Panama. For him, this was a 'particularly interesting region' for his customers, ironically referred to as 'La Ventosa' ('The Windy Place') – see figure 8. **DR23: Consider investigating ways of improving the reliability of underlying prediction data so that more predictions are made, skill scores become higher and therefore more trust can be placed in the predictions.**

Figure 8: 'La Ventosa' geographical region. Ironically, skill levels were insufficient to display predictions for 'The Windy Place'

Installed wind power bar

In a nutshell: The installed wind power bar was well-understood in general, but participants were unsure whether it related to capacity or power generated. All participants recommended changing the display unit from KW to MW.

The 'installed wind power' bar on the pop-up console menu was well-understood by participants in general, but there was some confusion – particularly on whether the installed wind power figure referred to capacity or power generated:

- P4 highlighted that it was unclear whether the installed wind power figure referred to capacity, or power generated. This confusion arose from reading the skill tooltip, which stated 'this chart shows how much *power is typically generated* from wind energy in the selected region.' **DR24: Consider being more consistent with interface descriptions of installed wind power.**
- P4 suggested that if displaying power generated, the interface should be clear about what period the power generated figure is referring to *"is it power generated in one week? One month?" One year?* (P4). Although he stated that knowing both capacity installed and power generated would be useful, he noted that it would be more difficult to calculate power generated – this figure may need to be estimated rather than based on observations. **DR24: Consider being more consistent with interface descriptions of installed wind power.**
- Similarly P5 *stated "this is not really understandable. What do you mean by 'typically generated'? Does it mean the power installed? If you're talking about capacity installed the measure will be kilowatts or even megawatts, but if you say capacity generated from wind generated then the figure would be kilowatts or megawatts per hour – which is the useful figure for wind energy. So I don't really understand what you mean by that definition… because it's contradictory… here you have 'installed wind power' and in the tooltip you talk about power generated"* (P5). **DR24: Consider being more consistent with interface descriptions of installed wind power. DR25: Consider displaying installed wind power in MW rather than KW.**

There was also confusion about why the 'skill' bar appeared to 'move' (was highlighted to indicate the skill percentage), while the 'installed wind power' did not:

- P2 asked *"why is this bar not moving?"* He was referring to the fact that while the 'skill' bar visually indicated the percentage skill level by filling some of the bar in light grey, the 'installed wind power' bar did not (the bar is always blue in colour, even when a wind farm has been selected and an amount of installed wind power is indicated (see figure 9). **DR26: Consider ensuring the 'installed wind power' bar is highlighted to show whether the installed wind power figure is relatively large or small.**

Figure 9: While the skill bar (top) is partially filled in grey to indicate 24%, the installed wind power bar (bottom) is not, even though there is over 1.5 Gigawatts of installed wind power in this area.

- P3 selected a prediction in an area of the map with no installed wind power. When asked why the installed wind power figure on the console was '0 KW,' he replied *"I have no idea!"* He then clicked on an area of the map with some installed wind power and the 'installed wind power' figure on the console changed to 645,000 KW. But P3 did not comment on the change, perhaps not noticing. **DR26: Consider ensuring the 'installed wind power' bar is highlighted to show whether the installed wind power figure is relatively large or small.**

All participants recommended changing the display unit of the 'installed wind power' bar from Kilowatts to Megawatts:

- When asked about the installed wind power figure (displayed on the interface as '12,000 KW'), P1 referred to the figure as '12 Megawatts' and commented that installed wind power is usually displayed in megawatts, *"because if I go here [clicks on another wind farm with an installed wind power figure of 597,020 KW], it starts to be a huge figure. Those 3 figures [597] would be enough."* Similarly P2 commented *"this number should be in Megawatts instead of Kilowatts."* He stated this was because if the visualisation was showing installed wind power in a particular area, the level of wind power installed would always be likely to be several thousand Kilowatts – making Megawatts a more useful unit to display this data in. P5 made a similar suggestion to refer to installed wind capacity in Megawatts. **DR25: Consider displaying installed wind power in MW rather than KW.**
- P3 referred to the installed wind power figure of '645,000 KW' as '645 Megawatts,' supporting P1's suggestion for displaying this figure in Megawatts. P3 stated that in the US, *"wind power plants are huge… typically they're displayed in Megawatts"* (P3). Similarly, P4 commented that, if referring to capacity of an installed wind farm, the unit of measure should be MW (rather than KW). However, if referring to power generated as the tooltip implies, the unit of measure should be MW/hr (P4's preferred option), or KW/hr, GW/hr). **DR25: Consider displaying installed wind power in MW rather than KW.**

Suggested new console

In a nutshell: We present an illustrative suggested new console aimed at complementing the suggested new legend in figure 4.

A potential new console is presented in figure 10. This console aims to complement the suggested new legend in figure 4. Again, as with the suggested new legend, this re-design is intended to be illustrative rather than prescriptive and there is scope to change the fine

details. In figure 10, the legend icons for 'skill' and 'installed wind power' have been duplicated to help users create a mental association between legend and console items. The question mark icons (for tooltips) have been enlarged and made a lighter shade of grey for greater contrast (addressing DR42, discussed later). The asterisk (*) icon has been replaced with a box drawn around the most likely tercile for consistency with the boxes used on the legend items in figure 4. This *partly* acts as an 'overall prediction' (see DR21), but it may be useful to carefully consider whether and how best to include an 'overall prediction score' in the predictions graph to support users in making decisions based on the predictions. The skill bar has been filled with the colour from the skill legend that corresponds to the skill percentage indicated in the bar. 'Installed wind power' is now displayed in Megawatts rather than Kilowatts. A caption has been added below the 'predictions' heading to explain that the predictions graph shows predicted wind speed in metres per second. The average now says 'median or mean' as a decision will need to be made on whether to switch to mean averages or keep medians (DR19).

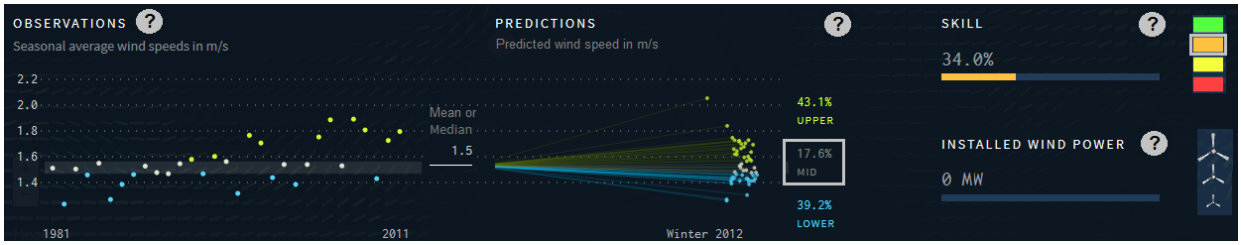

Figure 10: Suggested new console to complement suggested new legend.

Informational 'tooltips'

> ***In a nutshell: The tooltips were regarded as useful and informative, but not discovered by all participants. For some screen resolutions, they appear off-screen.***

The informational 'tooltips' were regarded as useful and informative by those participants who were aware of them and made use of them:

- P1 made extensive use of the tooltips, without instruction and commented that they were *"pretty well-made"* (P1). He used the tooltips for *"checking my guess"* of what aspects of the graphs meant.
- P2 read the predictions tooltip and thought it was 'clear.'

However, not all participants discovered the tooltips when exploring the interface in the context of their realistic information tasks; some only discovered them when prompted by the researcher:

- P5 eventually discovered the tooltips (when the researcher asked 'is there anything on the interface that could help explain' aspects of the console). **DR42: Consider enlarging the tooltip icon and making them a lighter shade of grey for greater contrast with the background.** Also, the right-end of the tip box did not appear on the screen (see figure 11). This only seems to happen in some screen resolutions but not others. **DR26: Consider investigating why tooltip text appears off-screen for some screen resolutions.**

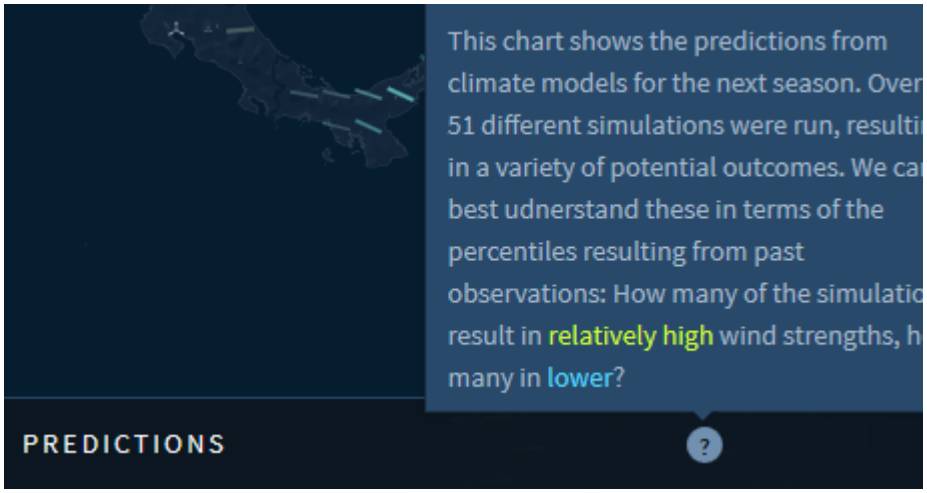

Figure 11: Typo in 'predicted change' tooltip text on the console and right-hand end of text box does not appear on the screen in certain resolutions

- P3 also eventually discovered the tooltips, but only when prompted by the researcher towards the end of the session. After reading them, he commented that they were *"very, very useful. I think it's very well-explained"* (P3). However, he still remained confused about the difference between predicted strength and predicted change after reading all the tooltips. **DR13: Consider including a caption on the legend to explain 'predicted change.'**

There is also a typo in the 'predicted change' tooltip text on the console:

- **DR27: Consider changing 'udnerstand' to 'understand' in the 'predicted change' tooltip text.** (See figure 11).

Settings menu

In a nutshell: The settings menu was only used by a couple of participants and may be superfluous.

One participant selected the 'skill and strength' option on the settings menu (figure 12) and thought 'not much' had changed. He suggested sticking to showing skill, strength and trend:

- P1 clicked on 'skill and strength' on the settings menu. When asked what changed on the interface when he did so, he stated *"not much. It's more colourful and all the rectangles are horizontal"* (P1).
- He stated *"I was actually wondering why you have those 2 options. I don't know… I would just stick to skill, strength and trend. I don't think I would use this second option because it doesn't bring me any more information. If I can see all the information at once, I have no reason to see less"* (P1). **DR28: Consider always displaying skill, strength and trend (rather than offering an option to display only skill and strength).**

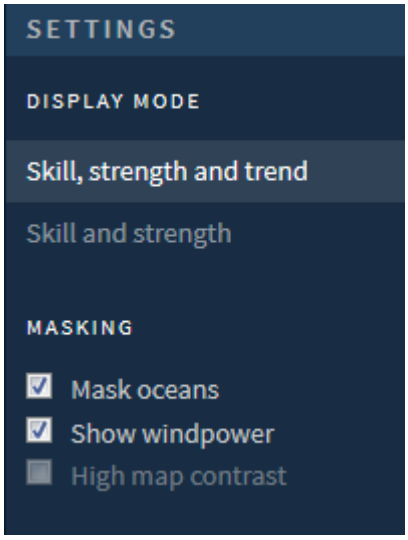

Figure 12: The settings menu options

The settings menu may actually be superfluous, as useful options might be either integrated into the map screen or used as default. For example:

- **DR29: Consider always displaying in high map contrast mode (either with a black or dark blue background)**, as derived from P5.
- **DR30: Consider moving the 'mask oceans' tickbox to the bottom of the legend to make the option to hide predictions for ocean areas more prominent** (as P4 suggested the need to be able to view only predictions 'on land' but did not interact with the settings menu).
- **DR31: Consider renaming the 'mask oceans' option to 'hide icons over oceans' to improve potential understandability**.

Zoom functionality

> ***In a nutshell: Several participants did not notice the zoom buttons right away because they are obscured with the console displaying the observation and prediction graphs.***

Several participants did not notice the zoom buttons right away. This was often because the zoom buttons were obscured much of the time with the console displaying the observation and predictions graphs. For example:

- P3 pointed his cursor to part of the map and asked *"can I zoom here?"* **DR32: Consider moving the zoom buttons to the top-left-hand-corner of the screen (under the 'Project Ukko' logo) and making them bigger**, as illustrated in figure 13.
- P2 commented *"I cannot zoom out with this [points cursor at console"* and P5 commented *"I can't see how to zoom in right now*" when he had the console window open (but he noticed the zoom button as soon as he closed the console window). Moving the zoom buttons should address this.

Figure 13: Larger zoom buttons, moved to below the 'Project Ukko' logo

- P2 asked *"you cannot zoom out totally. Why? This is not natural because we're used to working with Google Maps."* He would want to be able to zoom out further. We do not suggest this as a design recommendation as allowing users to zoom out further might make the icons on the map difficult to interpret unless they were dynamically re-clustered as users zoomed in and out.

Responsiveness

In a nutshell: There are minor responsiveness issues in version 2.1.0. Consider testing all future versions for responsiveness.

None of the participants who used version 2.1.0. of the prototype spontaneously mentioned the responsiveness of the tool, which indicates that there are no major responsiveness problems. However, when asked, P1 commented that there were minor responsiveness issues:

- When asked how responsive he thought the tool was when he zoomed in and out or clicked on something, P1 commented: *"you have a delay between the map layer and the placemark layer. That could certainly be improved, but it's nothing dramatic either."* **DR33: Consider improving map panning and zooming responsiveness.**

P5 happened to use version 2.2.0. of the prototype, as he had it previously bookmarked. This had more significant responsiveness issues than version 2.1.0:

- P5 struggled with map responsiveness when using version 2.2.0. of the prototype; when zoomed out fully in Google Chrome the map was slow to respond. It took several seconds to update when the participant scrolled. When he switched to version 2.2.1. the responsiveness improved slightly, but there was still a noticeable lag when scrolling. This participant was using a touchscreen device and he noticed that there was far less lag when swiping the screen with his finger than when scrolling with the mouse. **DR34: Consider investigating potential responsiveness issues specific to touchscreen/laptop hybrid devices.**

Additional desired functionality

In a nutshell: Several participants suggested additional functionality that they thought would make the prototype more useful to them. This was functionality beyond the improvements discussed already.

Several participants suggested additional functionality that the EUPORIAS team might want to consider incorporating into this or future versions of the prototype. While we have discussed some of the functionality they identified earlier in the report (where it has implications for improving parts of the interface they found confusing), participants also identified additional desired functionality, which we discuss here. It may be useful to consider whether adding some of these features in the future might make the seasonal wind forecast tool holistically more useful:

- P1 commented *"what I'm certainly missing here is something like an **export function**."* He commented he would want data from the graphs to be exported into a PDF, image format or Excel spreadsheet so he could present the data to others. This was supported by P5, who stated *"you only have a visual aspect of the data. I would need to have that data so I can compute different statistics with it"* (P5). For example, when looking at the 'observations' graph he suggested the need to create his own time series of the observation data to check how different the prediction is compared to the previous observation figures. He suggested the data was exportable in ASCII format. **DR35: Consider allowing users to export data.**
- P1 suggested the addition of functionality to **select multiple icons on the map at once** to receive prediction observation, prediction and skill data in the console related to averages for all the selected icons: *"What if I could draw a square around this particular region… and get an aggregate result of all the points contained in this region?"* (P1). Similarly P5 voiced the same need – to be able to *"select wind farms in batch mode; to get the info for several wind farms in a row."* (P5). P1 stated this could help a grid operator to prevent network congestion. He stated it could also help an inter-country energy trader to help them make decisions about what to buy in a particular region or country - *"providing a market trend indicator of how much wind power will be produced next season"* (P1). **DR36: Consider allowing users to view prediction data for multiple icon selections.**
- P2 suggested **labelling states and other regional boundaries** on the map. He struggled to find Texas, a known 'interesting area,' on the map of the USA: *"For big countries regional borders should be included"* (P2). P3 also struggled to find Texas state on the map, stating *"it's around here, right?"* **DR37: Consider labelling states and other regional boundaries on the map.**
- P3 wanted to be able to **set up different areas of interest** in North America and to be able to **review wind forecasts for this area on a regular basis**. **DR38: Consider allowing users to set up custom areas of interest and automatically providing them with new prediction data as it becomes available.**
- P3 mentioned the possibility of allowing users (e.g. wind farm owners) to **submit real observation data** with the purpose of enhancing the reliability of predictions. Similarly P5 suggested allowing the observation data on the prototype to be supplemented by past performance of his company's wind farm. **DR39: Allow users to submit real observation data to be used to enhance the predictive power of the tool.**
- P5 suggested the importance of updating the observation and prediction data every season. **DR40: Keep observation and prediction data as up-to-date as possible.**

Usefulness

In a nutshell: Participants found it difficult to answer how useful they found the prototype, as they found it hard to separate the prototype (which they generally thought was good) from the seasonal prediction data (which they generally thought was poor).

When asked to rate their perceived usefulness of the prototype on a scale of 0 to 4 (where 0 was 'not at all' and 4 was 'very'), participants found it difficult to answer. P1 did not provide a score. P2 rated it 3, P3 rated it 4 (but with provisos based on the usefulness of the data) and P5 between 1 and 2. The difficulties participants had in answering the question arose from them finding it hard to separate the prototype (which they generally thought was good) from the seasonal prediction data (which they generally thought was poor), as illustrated below:

- P1 highlighted that it is difficult to separate the prediction data provided by the prototype and the prototype itself: *"it's a complex question because there is the prototype, which is the interface and then there are models behind this… right now there is no evidence that [seasonal forecast data] is useful"* (P1). Models would need to provide better predictions in order to achieve this. P1 suggested this could be achieved by incorporating more observations, applying machine learning logic over initial predictions and observations. This highlights the importance of ensuring that prediction data is as accurate as possible so as to ensure the prototype is as useful as possible. P1 could not, however, think of any specific ways that the prototype itself could be changed to make it more useful.
- P3 also stated that the usefulness of the prototype was difficult to determine because *"we still have to validate the prediction model"* (P3). He also noted that its usefulness would depend on the availability of reliable data in the geographical areas of interest to the user. P4 made a similar comment, stating that the prototype will be more useful when uncertainty in the prediction models is reduced.

Usability

In a nutshell: Participants thought the prototype was, on average, 'very' easy to use and several used the word 'intuitive' to describe its usability.

When asked to rate the usability of the prototype on a scale of 0 to 4 (where 0 was 'not at all' and 4 was 'very'), P2 and P3 rated it a 4, P5 'between 3 and 4' and P1 and P4 rated it a 3. This makes an average of around 3.5 – 'very' easy to use, as supported by these comments:

- *"It's pretty well documented and quite intuitive. I didn't have to search for a lot to do such and such thing"* (P1).
- P2 thought the prototype was 'very easy' to use, 'well-designed' and 'well-coded.'
- P3 stated he thought the interface is 'quite good' and 'quite intuitive.'

P1 commented that he thought the prototype was easy to use despite the complexity of the seasonal wind forecasting domain:

- *"We're not talking about a type of science that's super-intuitive… I'd already seen these kinds of graphs…but I think the problem of seasonal forecasting is not necessarily very intuitive for common people or even for people from the industry"* (P1).

P4 encountered a bug in Internet Explorer where he would click on a particular icon on the map, but a different icon would be selected.

- For example, in figure 14, P4 clicked on the prediction rectangle and wind farm icon near the 'hand' mouse cursor, but an icon elsewhere on the screen (denoted by the purple shaded circle) was selected. **DR41: Continue to test for bugs in latest and previous versions of all major Web browsers.**

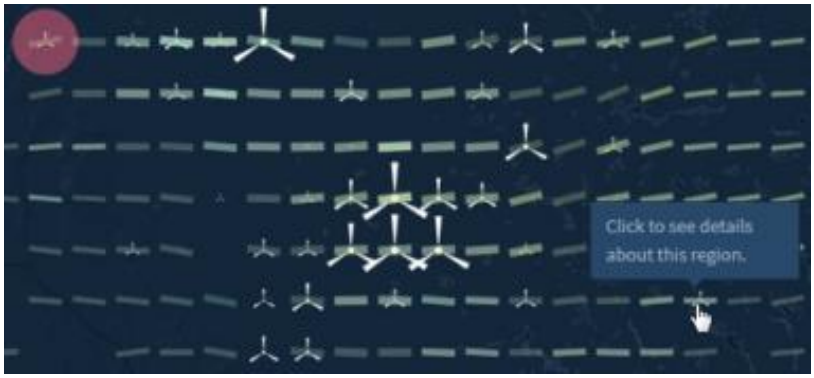

Figure 14: In Internet Explorer, clicking on one icon can select another

Likeliness of use

> ***In a nutshell: Participants stated they would only use the seasonal wind forecasting tool if and when seasonal wind prediction data becomes more reliable.***

Participants were very reluctant to rate their likeliness of use of the prototype; only P4 provided a tentative rating (of '0 if the quality of the seasonal prediction data is high and 4 if it is low'). Indeed, participants stated they were only likely to use the seasonal wind forecasting tool if and when seasonal wind prediction data becomes more reliable:

- "*It essentially depends on the regional skill. If the regional skill is over 10% I might look at it, 20% I'm likely to and 30% I will certainly"* (P1).
- Similarly P3 commented *"it depends how trustful it demonstrates it is… we still have to demonstrate confidence in the prediction models"* (P3) and P5 commented that *"this is a tough one because first we have to know how useful the predictions are. It's not about the prototype, it's about the predictions. The problem is not really about the design of the prototype…*" (P5).
- P5 commented that the useful seasonal predictions for his company are likely to be where they already have wind farms. But because the skill level is too low in many of these areas, no prediction data is displayed and this impacts on his company's likeliness of use of the prototype.
- P2 stated *"I'm sorry to be pessimistic, but they will not use this until the predictability is higher…The main focus for all of these projects should actually be on how predictability can be used to make decisions."* (P2). He explained further that in his opinion *"increasing the predictability is in the hands of the modellers, but translating this skill into something that pushes people to make decisions is not in the hands of these people. It's in the hands of other people, who I don't know who they are"* (P2).

The above comments are all related to **DR23: Consider investigating ways of improving the reliability of underlying prediction data so that more predictions are made, skill scores become higher and therefore more trust can be placed in the predictions.**

One participant also commented on the importance of being able to make use of the prediction data to make decisions:

- P4 commented that if the prototype was not able to provide him with clear overall predictions, he would be unlikely to use the prototype. He gave an example of a selected region with 31.4% of predictions in the upper tercile, 35.3% in the mid and 33.3% in the lower tercile (see figure 15). P4 thought this 'split prediction' for the

region was not useful. **DR21: Consider providing an overall interpretation of the prediction data – e.g. an overall prediction score and some on-screen guidance (e.g. in a tooltip) about how to interpret it.**

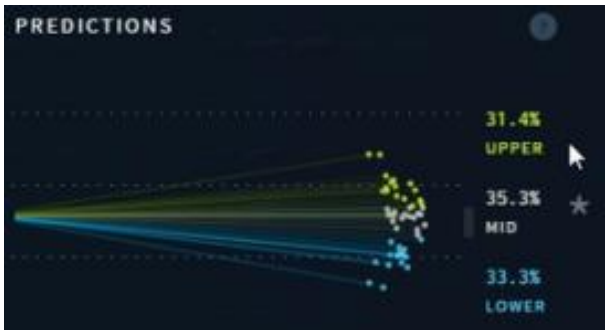

Figure 15: A 'split prediction' with similar percentages for the upper, mid and lower terciles

Summary of design recommendations and suggested priorities

We finish by summarising our design recommendations and by making suggestions for how we would priorities them. We do not recommend that EUPORIAS tries to implement all these recommendations (in fact, we have suggested that several are only implemented after consultation with several additional potential users). It is also important to note that we have been inclusive with our design recommendations; even if only one participant has experienced an issue with the interface or recommended additional functionality, we have included it as a potential recommendation. We would recommend first implementing those design recommendations you consider to be **very important** and **very easy to implement**. (V, V in table 1 and shaded in the darkest shade of grey). Next, we recommend you implementing those recommendations you consider to be **very important** but **not so easy to implement** (V, N in table 1 and shaded in next the darkest shade of grey), followed by those that you consider **not so important** but **very easy to implement** (N, V in table 1). Your final priority should be those you consider to be **not so important** and **not so easy to implement** (N, N in table 1 and shaded the lightest shade of grey). Note that these are only indicative suggestions of priorities – you are likely to be best placed to make more tailored and accurate priority judgements.

Design recommendation	**Potential importance (Very, Not Very)**	**Potential ease of implementing (Very, Not Very)**
DR1: Consider including informational 'tooltips' for legend items as well as for items in the popup console window	V	V
DR2: Consider whether or not to display prediction data where wind farms currently exist and the skill level is negative.	V	V
DR3: Consider using different rectangle colours to denote difference in skill level, different rectangle thicknesses to denote difference in predicted strength and different sloping angles to denote predicted change.	V	V

DR4: Consider renaming 'predicted strength' to 'predicted wind speed.'	V	V
DR5: Consider changing the bandings for different rectangle thicknesses for 'predicted strength,' to display a greater variety of rectangle thicknesses on the map.	N	V
DR6: Consider explaining on the interface what 'seasonal' means and which months correspond to each season.	N	V
DR7: Consider whether or not it is appropriate to rename 'predicted change' legend item to 'next winder trend' or similar.	N	V
DR8: Consider renaming 'installed wind power' to 'installed wind capacity' or something similar. We suggest 'wind power installed in region.'	V	V
DR9: Explain in an informational tooltip for 'installed wind power' that the icons represent wind capacity in a specific geographical region, *not* individual wind farms.	V	V
DR10: Consider using only 3 different sizes of 'installed wind power' icon on the map.	V	V
DR11: Consider whether it is necessary to include predictions in areas where there are no existing or prospective wind farms installed.	N	N
DR12: Consider whether or not to show individual existing wind farm installations and possibly forecast installations on the map.	N	N
DR13: Consider including a caption on the legend to explain 'predicted change.'	V	V
DR14: Consider highlighting the relevant legend item icon to indicate how the selected map icon relates to the legend.	V	V
DR15: Consider explaining in the observations graph tooltip whether the graph uses observations data from successive same seasons to come up with predictions for that same season in the coming year.	V	V
DR16: Consider giving users the option to change the season for which the predictions are for (currently always 'Winter 2012').	V	V
DR17: Consider a line graph view instead of or in addition to a scatterplot view.	N	N
DR18: Consider mentioning that the intervals are terciles in the 'predictions' informational tooltip.	N	V
DR19: Consider either using mean instead of median to denote average in the predictions graph, or explain in the 'predictions' tooltip why median instead of mean is used.	N	V
DR20: Consider explaining in the 'observations' tooltip that the median is calculated of all 51 observations on the observations graph.	N	V

DR21: Consider providing an overall interpretation of the prediction data – e.g. an overall prediction score and some on-screen guidance (e.g. in a tooltip) about how to interpret it.	V	N
DR22: Consider providing a scale for each legend item.	V	V
DR23: Consider investigating ways of improving the reliability of underlying prediction data.	V	N
DR24: Consider being more consistent with interface descriptions of installed wind power.	V	V
DR25: Consider displaying installed wind power in MW rather than KW.	N	V
DR26: Consider investigating why tooltip text appears off-screen for some screen resolutions.	N	V
DR27: Consider changing 'udnerstand' to 'understand' in the 'predicted change' tooltip text.	N	V
DR28: Consider always displaying skill, strength and trend (rather than offering an option to display only skill and strength).	N	V
DR29: Consider always displaying in high map contrast mode (either with a black or dark blue background).	N	V
DR30: Consider moving the 'mask oceans' tickbox to the bottom of the legend to make the option to hide predictions for ocean areas more prominent.	N	V
DR31: Consider renaming the 'mask oceans' option to 'hide icons over oceans' to improve potential understandability.	N	V
DR32: Consider moving the zoom buttons to the top-left-hand-corner of the screen (under the 'Project Ukko' logo) and making them bigger	V	V
DR33: Consider improving map panning and zooming responsiveness.	N	N
DR34: Consider investigating potential responsiveness issues specific to touchscreen/laptop hybrid devices.	N	N
DR35: Consider allowing users to export data.	N	N
DR36: Consider allowing users to view prediction data for multiple icon selections.	N	N
DR37: Consider labelling states and other regional boundaries on the map.	V	N
DR38: Consider allowing users to set up custom areas of interest and automatically providing them with new prediction data as it becomes available.	N	N
DR39: Allow users to submit real observation data to be used to enhance the predictive power of the tool.	N	N
DR40: Keep observation and prediction data as up-to-date as possible.	V	N
DR41: Continue to test for bugs in latest and previous versions of all major Web browsers.	V	V

Table 1: Summary of design recommendations with an illustrative indication of potential importance and potential ease of implementing.

References

Blandford, A., Green, T., Furniss, D., Makri, S. (2008). Evaluating system utility and conceptual fit using CASSM. *International Journal of Human Computer Studies* 66(6), 393-409.

Card, S. K., & Moran, T. P. (1988). User technology: From pointing to pondering. In *A history of personal workstations* (489-526). ACM.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Makri, S., Blandford, A., & Cox, A. L. (2011). This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behaviour. *Information Processing & Management*, *47*(3), 336-348.

Makri, S., Blandford, A., Gow, J., Rimmer, J., Warwick, C., & Buchanan, G. (2007). A library or just another information resource? A case study of users' mental models of traditional and digital libraries. *Journal of the American Society for Information Science and Technology*, *58*(3), 433-445.

Munzner, T. (2009). A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on*, *15*(6), 921-928.

Norman, D. A. (1983). Some observations on mental models. *Mental models*, *7*(112), 7-14.

Vaughan, C., & Dessai, S. (2014). Climate services for society: origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(5), 587-603.

Appendix A: Observation script and interview guide

Note: As the questions asked during the evaluation were opportunistic and probing, they are not all listed here. The listed questions provide an illustration of the types of questions asked and the breadth of topics covered.

Part 1: Explaining the research

"Thank you very much for agreeing to take part in this research. In it, I'll be asking you to use a prototype tool that aims to support people in using seasonal wind predictions to make decisions. I'm interested in learning about your understanding of the tool - its purpose and how it works. So I won't be demonstrating it beforehand. Instead, I'll ask you to use the tool and will ask you questions as you use it. At the end, I'll also ask you some questions about how useful and easy to use of think the tool is and how likely you are to use the tool in the future."

"Your participation in this research will not only allow us to determine the success of the prototype and to make improvements, but will also inform the development of future wind prediction tools."

- Is the **aim** of this research clear? Do you have any **questions** about the aim?

"The session will last between 30-45 mins. and I'll ask you to share your screen with me. I'll be recording the screen and audio. You're free to stop the study at any time and will not be penalised in any way if you do stop the study. You're welcome to review the recording after the session or ask us to delete it. Just send me an email if you'd like to do that."

- Is that okay? Do you have any **questions or concerns**?

Part 2: Introducing the prototype

[Researcher invites participant to read the text on the prototype Website homepage].

- Do you **have any questions** about what you have read or what the prototype is intended for?
- Have you **previously used** the prototype or earlier versions? When? What for?
- Have you **previously provided any feedback** on the prototype or earlier versions? If yes, what was it?

Part 3: Introductory questions

- Can you tell me a little about your **job role** and how using a prototype that shows **seasonal wind forecasts might help you in your job role**?
- Do you or your organisation currently use **seasonal wind forecasts**? If no, why not? If yes, what for? Can you give me an example?
- What different **types of tasks** would you like to perform using a seasonal wind forecasting tool and what **types of information** would you like a seasonal wind forecasting tool to provide you with? *[To address Objective 1: Domain situation and Objective 2: Task and data abstraction].*

Part 4: Participant to explore the tool in the context of a realistic task

Researcher to ask probing, opportunistic questions aimed at uncovering their understanding of the different 'layers' of the prototype (e.g. the interface and visualisation)].

"Could you now use the prototype to carry out a realistic task, perhaps one of the tasks you mentioned previously. I would like you to 'think aloud' while you are interacting with the prototype – telling me exactly what you're doing and why. For example 'I'm looking at this part of the screen, I'm clicking on this button.' I'll be asking you questions as you do that. If there's enough time, you can carry out more of your tasks using the prototype."

- Do you have any questions about what you have to do?

Note: No structured questions will be asked in this part of the evaluation, but some examples to tailor depending on participants' task, interface behaviour and actions may include (but are not limited to):

- What do the different icons on the legend (**skill**, **predicted strength**, **predicted change**, **installed wind power**) mean to you? Are there any ambiguities in the legend? If yes, what are they? How might we make the legend more useful? *[Addresses the special interests of interpretability of multi-dimensional line elements].*
- How does the prototype make use of **colour**? What do the different colours mean to you? Can we make better use of colours? If yes, how?
- When you click on a symbol on the map, a console pops up at the bottom of the screen. What do the different graphs (**observations** and **predictions**) mean to you? Can you explain your understanding of them to me? How useful are each of the graphs and individual components of them useful? What makes them useful/not useful? How might we make them more useful? *[Addresses the special interests of interpretability of the 'ray model,' presenting forecast tendency in terms of terciles of historical data].*
- What does the '**skill**' figure on the console mean? What constitutes a high/low skill level? How useful is knowing the skill level? Why? How could we make it (more) useful? *[Addresses the special interest of understandability of the concept of skill and its role in the visualisation].*
- What does the **'installed wind power**' figure on the console mean? What constitutes a high/low level of installed wind power? How useful is knowing the level of installed wind power? Why? How could we make it (more) useful?
- How useful are the **tips** that pop up for explaining elements of the visualisation? Did any of the tips help you better understand the visualisation? If so, which ones and how? How could we improve the tips?
- What happens when you select different options on the **'settings'** menu? Why do the icons on the map all become horizontal when 'skill and strength' rather than 'skill, strength and trend' is chosen from the settings menu? Are there any other settings you would like to be able to change? Why?
- How **responsive** is the tool when you zoom in and out or click on something on the interface? In your opinion, is it responsive enough? Why? *[Addresses the 'algorithm' part of Tamara Munzner's nested model].*

Part 5: Wrap-up questions focused on participants' perceptions of prototype success (in relation to success criteria)

On a scale of 0 to 4, where 0 is 'not at all' and 4 is 'very,' how...

- **Useful** do you think the prototype is? Why? How might we make it more useful?

- **Easy to use** do you think the prototype is? Why? How might we make it easier to use?
- **Likely are you to use** the prototype in the future? Why? How might we make it more likely that you will use the prototype in the future?

Thank you very much for your time. We really appreciate your help and hope you found using the prototype interesting.